A BOTTOM UP HYBRID METHOD FOR ISOLATED WORDS RECOGNITION

Olivier Delemar & Harouna Kabré
Institut de la Communication Parlée, Grenoble, France

ABSTRACT

In this article we present a new hybrid method for speech recognition which uses expert knowledge to control Hidden Markov Models in a purely bottom-up way. The phonetic knowledge used by the hybrid model to perform this control is embedded in the standard hidden Markov models by the way of an "expert matrix" which expresses whether a given broad phonetic label may or may not, be detected by the expert system while the model is in a given state.

INTRODUCTION

Many recent works in the field of speech recognition tend to combine different methods of acoustic-phonetic decoding, in order to take advantage of their particularities. Among the different hybrid methods, there are those which use the discrimination power of neural networks with the time alignment capacities of Hidden Markov Models [1] [2] [3] and those which combine rule based systems with HMM [4] or neural networks [5].

Although the Markov modeling is one of the best speech recognition methods, it still has great difficulty coping with explicit phonetic knowledge. A usual solution to this problem is to constrain HMM to assign one state either for one particular acoustic phase of the speech unit [6] [7] [8], or for one articulatory configuration of the vocal track [9]. The underlying principle of these methods is that they force the matching of the underlying phonetic structure - which may be described by a human expert - by the state transition graph of the Markovian process.

This paper presents a hybrid recognition method which uses expert rules to add automatically phonetic knowledge to the set of standard HMM parameters during the training step and to control the bottom-up recognition process according to that knowledge.

The rest of the paper is organized as follows: the first section will present the standard HMM principle, then the hybrid model will be described with an example showing how it controls the recognition algorithm. The last section will give some results and will discuss the enhancements and limitations of this method.

HIDDEN MARKOV MODELING

HMMs are finite states automatons which model sequences of quasi-stationary phases [10]. They are formed by a fixed number of states linked to others by arcs. Each arc has an associated transition probability - possibly null - while each state is associated with an emission probability density function (pdf). An N-states Markov model is then defined entirely by it's A-matrix of transition probabilities a_{ij} and it's set of emission pdf's $b_i(.)$ called the B-matrix.

Speech recognition with HMM consists in evaluating each model probability, given a vector of acoustic features. This may be performed by the Viterbi algorithm [11] which, in addition, finds the best path along the Markov chain in order to maximize this probability. During this process, the choice of a transition from state i to state j, given the feature vector O_t, depends on the value $a_{ij}*b(O_t)$ which may be seen as the "cost" of the transition from state i to state j, weighted by the "distance" between the j^{th} state's inner representation of an acoustic configuration and the observed acoustic feature O_t. Thus the Viterbi algorithm performs nothing other than a time alignment procedure. Figure 1 shows such an alignment for the French word "ouvre" (/uvR@/).

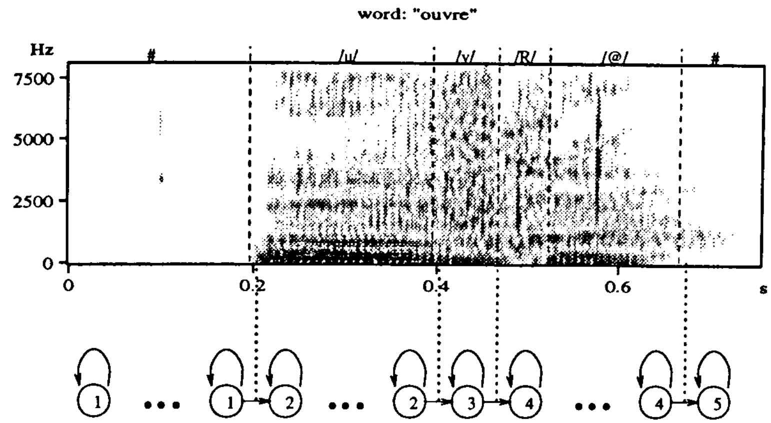

Figure 1: Time alignment of the word model "ouvre" with an occurrence of this word: the transitions from one state to the next one are quite exactly synchronized with the manually determined phonemes' frontiers.

The efficiency of the acoustic-decoding with HMM relies on the optimization of the models' parameters. The forward-backward algorithm [12] which is widely used for this purpose is an Expectation-Maximization procedure which iteratively re-estimates the transition and emission probability, given a training set of acoustic data. It should be noticed that, as far as HMMs are probabilistic models, the number and the quality of the training samples condition the representativity and the generalization capacity of the models. Furthermore, short acoustic events like the stop-consonants' burst are poorly modeled because of their reduced number of representative feature vectors.

THE HYBRID MODEL

The hybrid model presented here tries to satisfy three constraints: introducing phonological knowledge expressed by an expert system into the HMM's decision procedure, keeping the automatic aspect of the model's training phase and maintaining the bottom-up aspect of the acoustic-phonetic decoding by HMM which allows a real-time implementation for speech recognition. Taking into account the speech segmentation generated by the Viterbi algorithm, the training phase of the hybrid model will create a so called "expert matrix" E with as many rows as the Markovian model has states and as many columns as there are broad phonetic classes predicted by the expert system.

In our experiments, the expert system is a set of deterministic networks [13] finding occurrences of voiceless fricatives and stop consonants by applying fuzzy thresholds to the zero-crossing rate, the power and its first and second order derivative. Thus, after a standard model training, a time alignment is achieved for each training sample and compared to the broad phonetic labels in order to evaluate the "plausibility" that this label will be predicted by the expert system while the i^{th} state of the model is being visited.

During the recognition phase, both Markov models and the expert system are run simultaneously. Each time a label is generated by a deterministic network, the expert matrices are parsed to determine the states of each models which may be visited at this time. All other states are weighted so that any state sequence containing these states is forbidden. The constraints applied to HMM are then dynamic, and the hybrid model is still bottom-up. They are also time synchronous and the expert knowledge is then taken into account by the time alignment process.

RESULTS AND DISCUSSION

The hybrid model has been tested on a subset of the French database BDSON. This corpus is formed by 161 mono or bi-syllabic words, each pronounced once by 5 male and 5 female speakers. The speech signal is sampled at 16kHz and analyzed every 10ms by a Perceptual Linear Predictive Coding algorithm [14] to produce the feature vectors. Both standard and hybrid models are trained with 6 of the 10 utterances and the other 4 are used as the test set.

For this corpus, the standard HMMs give a 38% recognition rate (49% if we consider the first two candidates). By correcting 16% of the confused words, the hybrid model increases this rate to 43% (56% for the first two candidates).

word "autre" is not penalized because the best states sequence implies that the appropriate state is visited when the voiceless plosive is detected and this model becomes the most likely one.

As a result of the time synchronous constraints imposed on the models' states sequences, the hybrid model demonstrates on the ability to closely model the phonetic structure of words, even with a small number of training utterances. This is due to the fact that the a priori knowledge-based expert system is not dependent on the quantity of training data. Furthermore, the principle

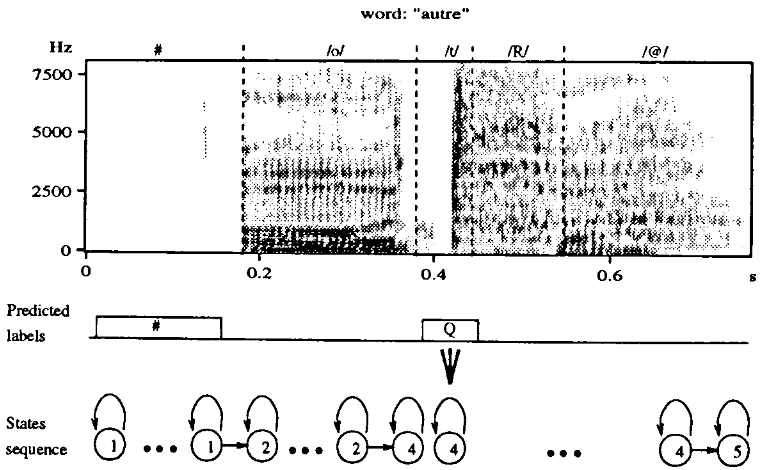

Figure 2: How the expert system controls the Markovian model: the path followed by the Viterbi algorithm along model's states is penalized when the "Q" label is predicted by the expert system.

An example of wrong-model correction is shown in figure 2. As a result of differences in pronounciation between the speaker and those used to train the models, the uttered word "autre" (/otR@/) happens to be more likely emitted by the model of the word "ouvre". On the other hand, acoustic events like the short silence followed by a noised burst are clearly detected by the expert system which then produces the label Q, indicating a voiceless plosive. As long as there is no state -according to the E matrix- which may be visited while this phonetic label occurs, all states' probability are weighted down, resulting in a reduction in the model's final probability. By contrast, the model of the

of time rendez-vous imposed on HMMs' states sequences by the external system may be extended to handle other types of information.

The principle of this method is that it uses the external information whenever it is available, this means that the expert system has to be as robust as possible. In fact, some unexpected detections made by the expert system cause the correct model to be penalized. Thus further works will tend to optimize the performance of the expert system in order to close approach the results obtained in a preliminary experiment, using manualy given labels obtained from a human expert [15].

CONCLUSION

We have described a new hybrid approach to speech recognition using HMMs and a rule-based expert system. Phonological knowledge as expressed by the expert system is embedded in the models by the way of the E matrix. This knowledge is then used during a purely bottom-up recognition process, to constrain the states sequence and avoid forbidden states. The results are encouraging but the rules have to be optimized in order to produce solely robust information.

REFERENCES

[1] Franzini M. A., Witbrock M. J. and Lee K-F (1989). "A Connectionist Approach to Continuous Speech Recognition", *ICASSP 89*, vol. 1, pp. 425-428.

[2] Bourlard H. and Wellekens C. J. (1990) "Links between Markov Models and Multilayer Perceptrons", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, no. 12, pp. 1167-1178.

[3] Renals S., Morgan N., Bourlard H., Cohen M. and Franco H. (1994). "Connectionist Probability Estimators in HMM Speech Recognition", *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 1, pp 161-174.

[4] Haton J. P., Carbonell N., Fohr D., Mari J. F. and Kriouille A. (1987) "Interaction between Stochastic Modeling and Knowledge-Based Techniques in Acoustic-Phonetic Decoding of Speech", *Proc. ICASSP*, vol. 2, pp. 868-871.

[5] Nocera P. and Bulot R. (1994) "Utilisation d'une méthode mixte : expertise et réseaux de neurones pour la reconnaissance des occlusives du français", *Reconnaissance automatique de la parole*, GDR-PRC Communication Homme-Machine.

[6] Deng L., Lenning M. and Mermelstein P. (1990) "Modeling Microsegments of Stop Consonants in a Hidden Markov Model Based Word Recognizer", *JASA 87(6)*, pp. 2738-2747.

[7] Farhat A., Pérennou G. and André-Obrecht R. (1993) "A Segmental Approach versus a Centiseconde one for Automatic Phonetic Time-Alignment", *EUROSPEECH 93*, vol. 1, pp. 657-660.

[8] Jouvet D., Barthova K. and Mouné J. (1991) "On the Modelization of Allophones in an HMM-Based Speech Recognition System", *EUROSPEECH 91*, vol. 2, pp 923-926.

[9] Deng L. and Erler K. (1992) "Structural Design of Hidden Markov Model Speech Recognizer using Multivalued Phonetic Features: Comparison with Segmental Speech Units", *JASA 92(6)*, pp. 3058-3067.

[10] Rabiner L. (1989). "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. IEEE*, vol. 77, pp. 257-285.

[11] Viterbi A. J. (1967). "Error Bounds for Convolution al Codes and an Asymptotically Optimal Decoding Algorithm", *IEEE Trans. Informat. Theory*, vol. IT-13, pp. 260-269.

[12] Baum L. E. (1972). "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes", *Inequalities*, vol. 3, pp. 1-8.

[13] Kabré H. (1991) "Décodage acoustico-phonétique multilingue : système à base de connaissances et étiquetage automatique de corpus de parole", *PhD Thesis*, Université Paul Sabatier, Toulouse.

[14] Hermansky H. (1990) "Perceptual Linear Predictive Coding of Speech", *JASA 87(4)*, pp. 1738-1752.

[15] Delemar O. (1994) "Reconnaissance de mots enchaînés par une méthode hybride : réseau markovien et base de règles", *Proc. 20èmes Journées d'Etude sur la Parole*, pp. 497-500.